UNECE

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

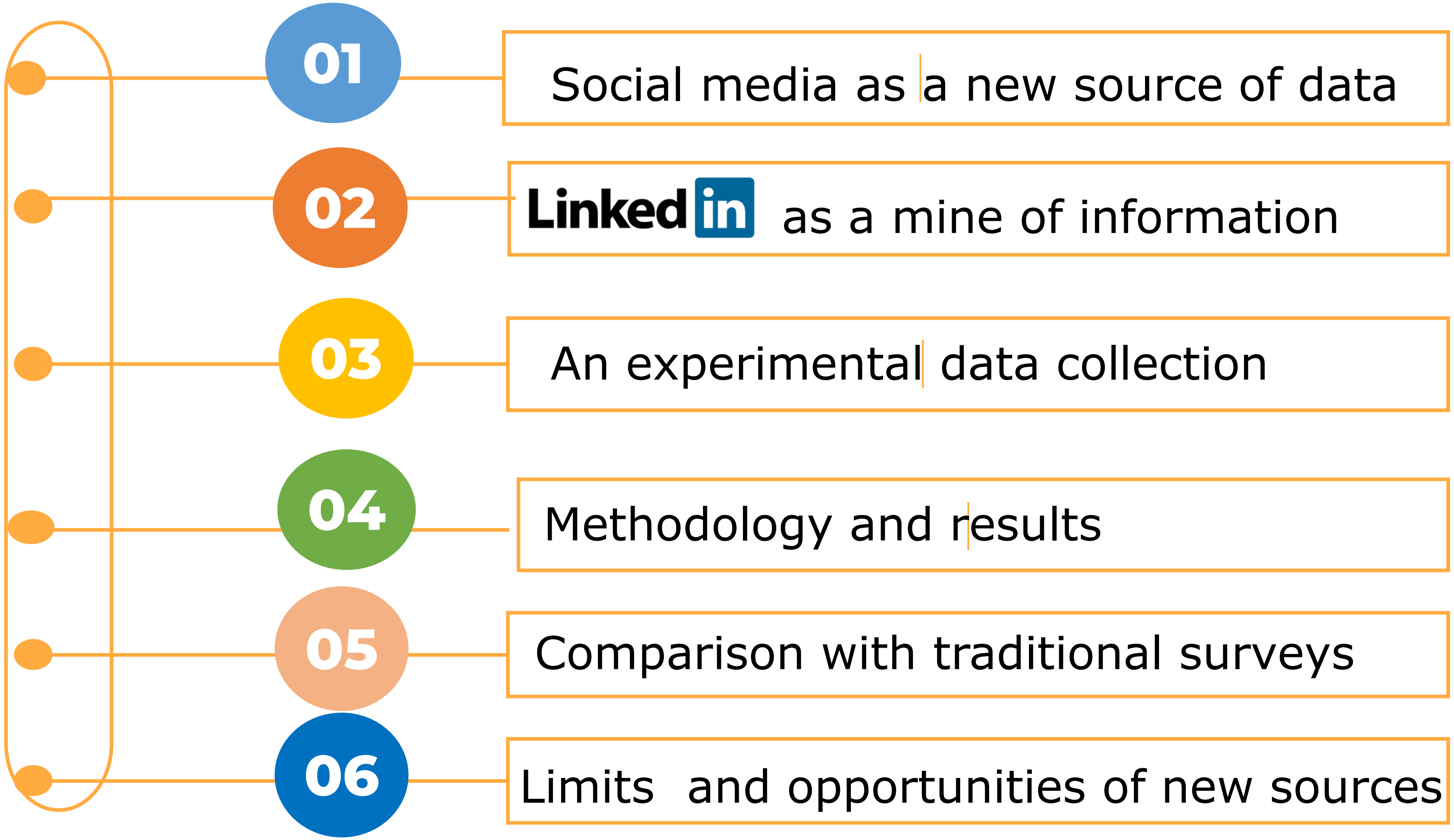
Expert Meeting on Statistical Data Collection and Sources
22-24 May 2024, Geneva, Switzerland

Use of A.I. to make LinkedIn as a new data source

Istat Istituto Nazionale di Statistica

Simona Cafieri (Italian National Institute of Statistics, Italy)

Outline



01 Social media as a new source of data

02 LinkedIn as a mine of information

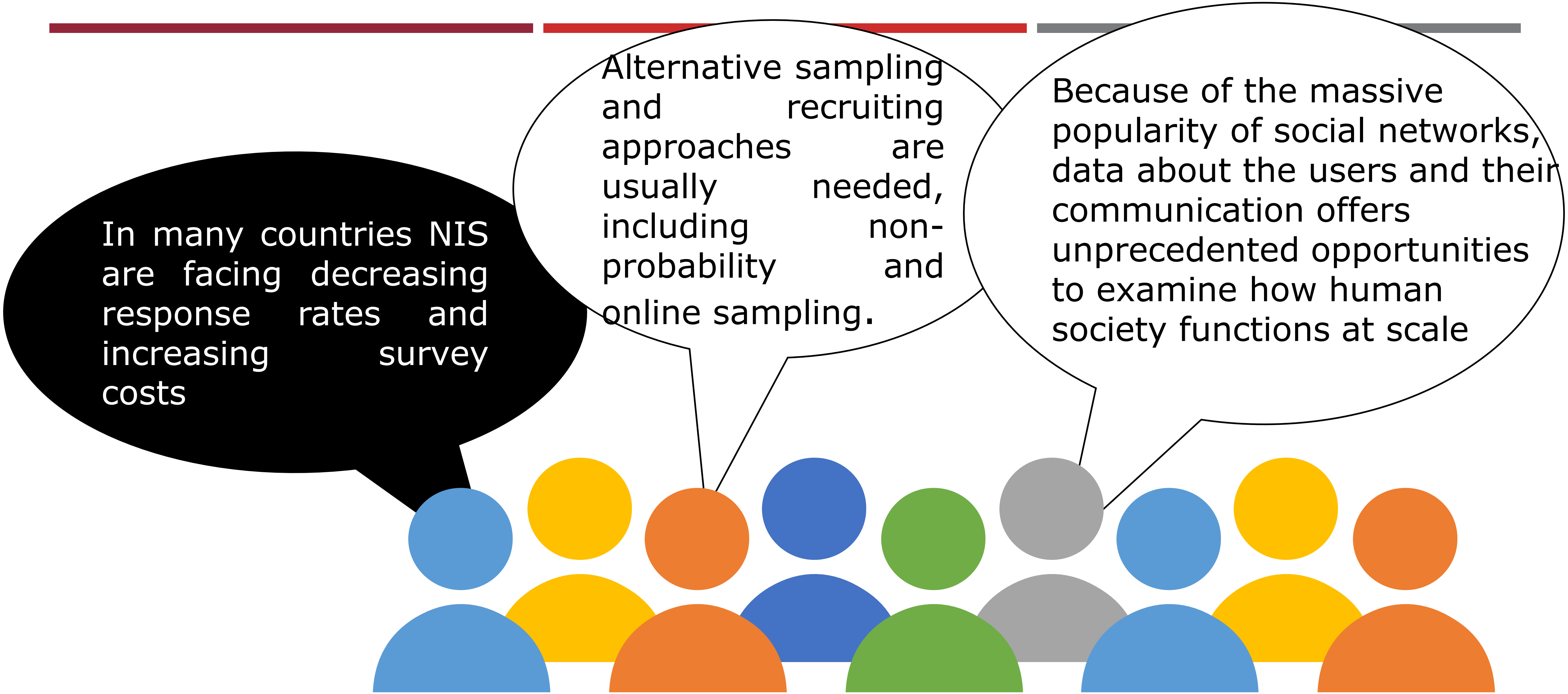
03 An experimental data collection

04 Methodology and results

05 Comparison with traditional surveys

06 Limits and opportunities of new sources

Background



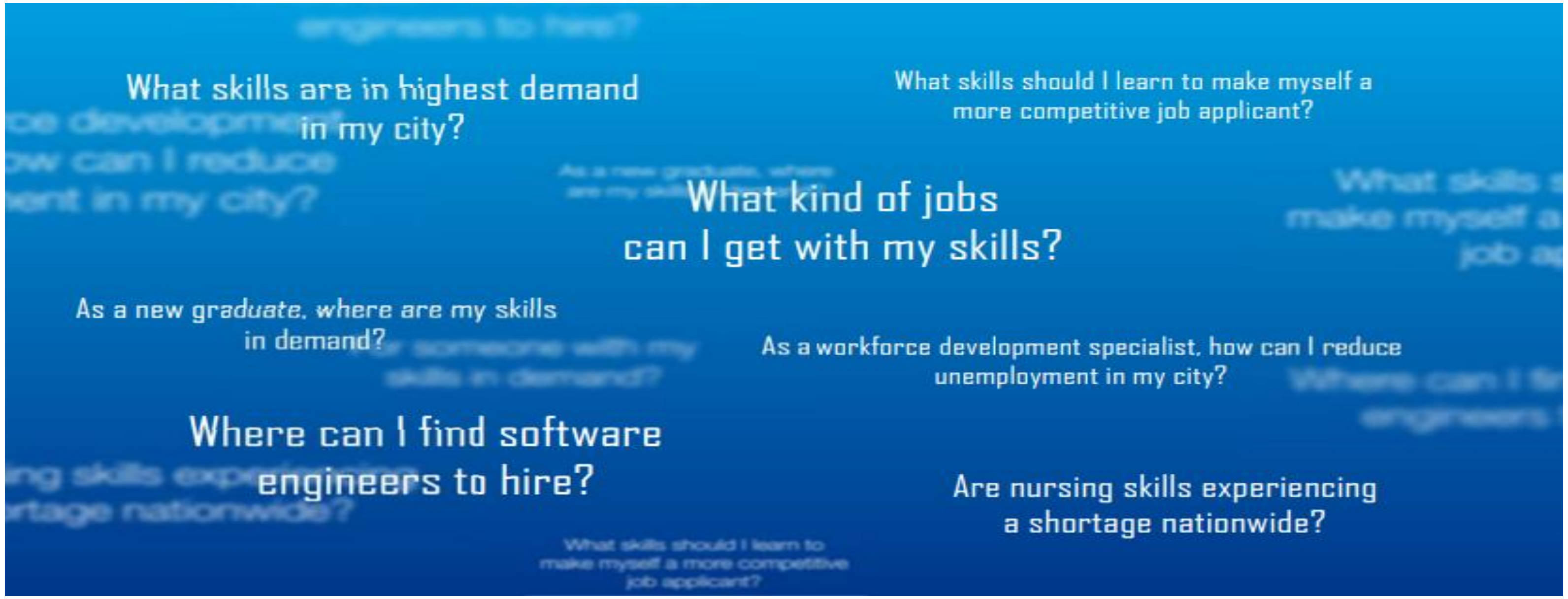
In many countries NIS are facing decreasing response rates and increasing survey costs

Alternative sampling and recruiting approaches are usually needed, including non-probability and online sampling.

Because of the massive popularity of social networks, data about the users and their communication offers unprecedented opportunities to examine how human society functions at scale

Social media as a new source of data

- ➢ They represent a growing portion of the general population
- ➢ Large amount of meta information available on these platforms
- ➢ Data collection is much cheaper than other sources
- ➢ We get data on populations that would be unlikely to respond to statistical surveys



What skills are in highest demand in my city?

What skills should I learn to make myself a more competitive job applicant?

What kind of jobs can I get with my skills?

As a new graduate, where are my skills in demand?

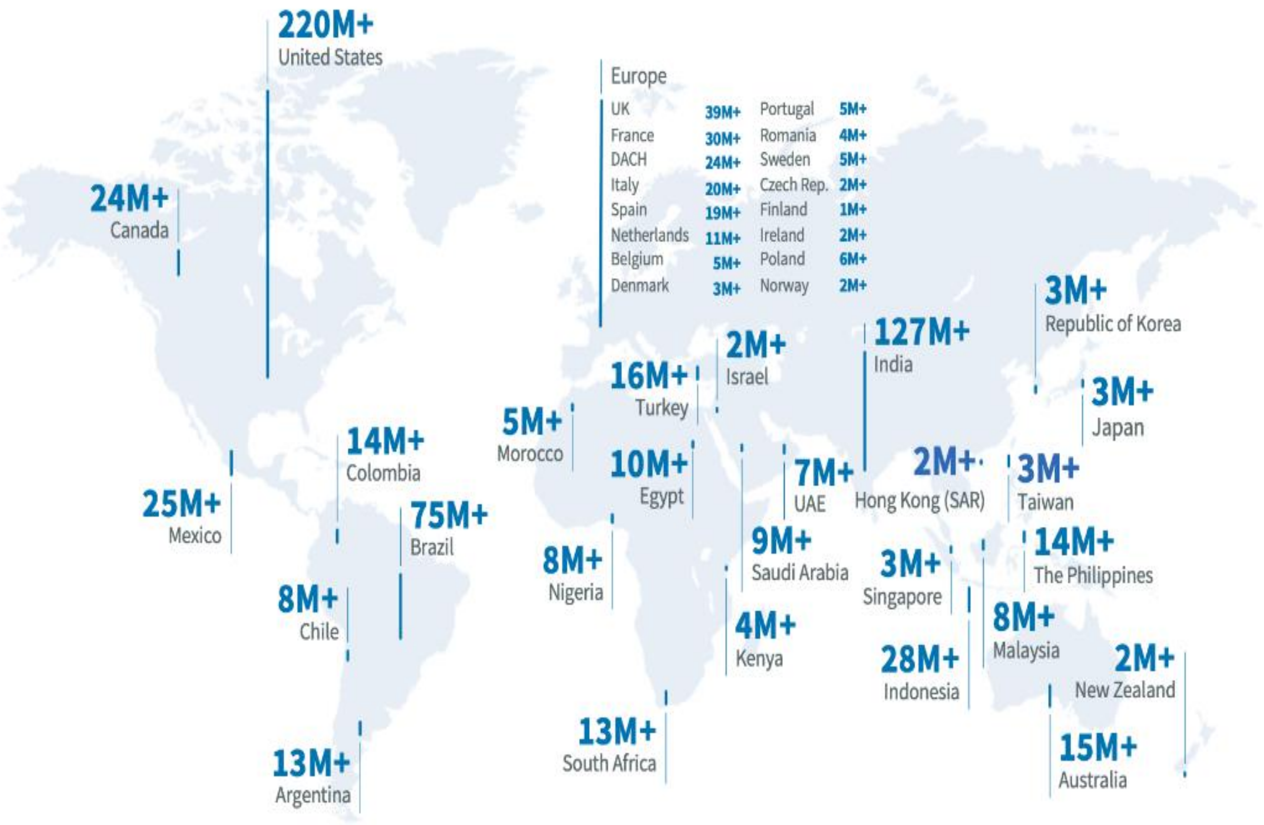
As a workforce development specialist, how can I reduce unemployment in my city?

Where can I find software engineers to hire?

Are nursing skills experiencing a shortage nationwide?

LinkedIn as a new source of data

Is the most comprehensive and the most up-to-date professional database,
a very mine of informations
(so it is a very precious source!)



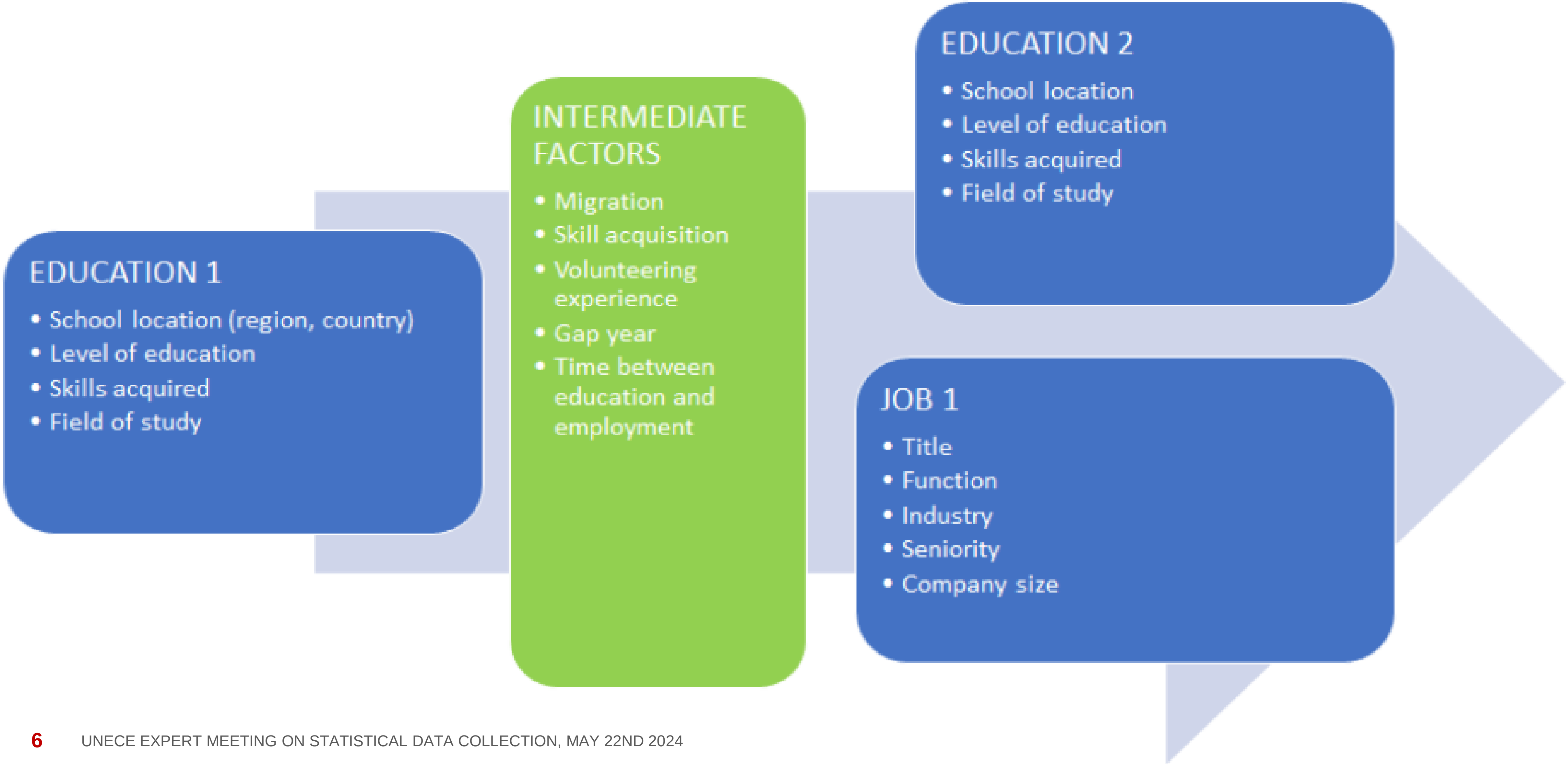
LinkedIn profiles: information about education, work experience, skills and interests of users

Job vacancies: information about position, requirements and qualifications

Company pages: information about business, products and services

Industry trends: data for market analysis and personnel acquisition strategies.

Generalised framework for LinkedIn research



EDUCATION 1

- School location (region, country)
- Level of education
- Skills acquired
- Field of study

INTERMEDIATE FACTORS

- Migration
- Skill acquisition
- Volunteering experience
- Gap year
- Time between education and employment

EDUCATION 2

- School location
- Level of education
- Skills acquired
- Field of study

JOB 1

- Title
- Function
- Industry
- Seniority
- Company size

LinkedIn as a new source of data:

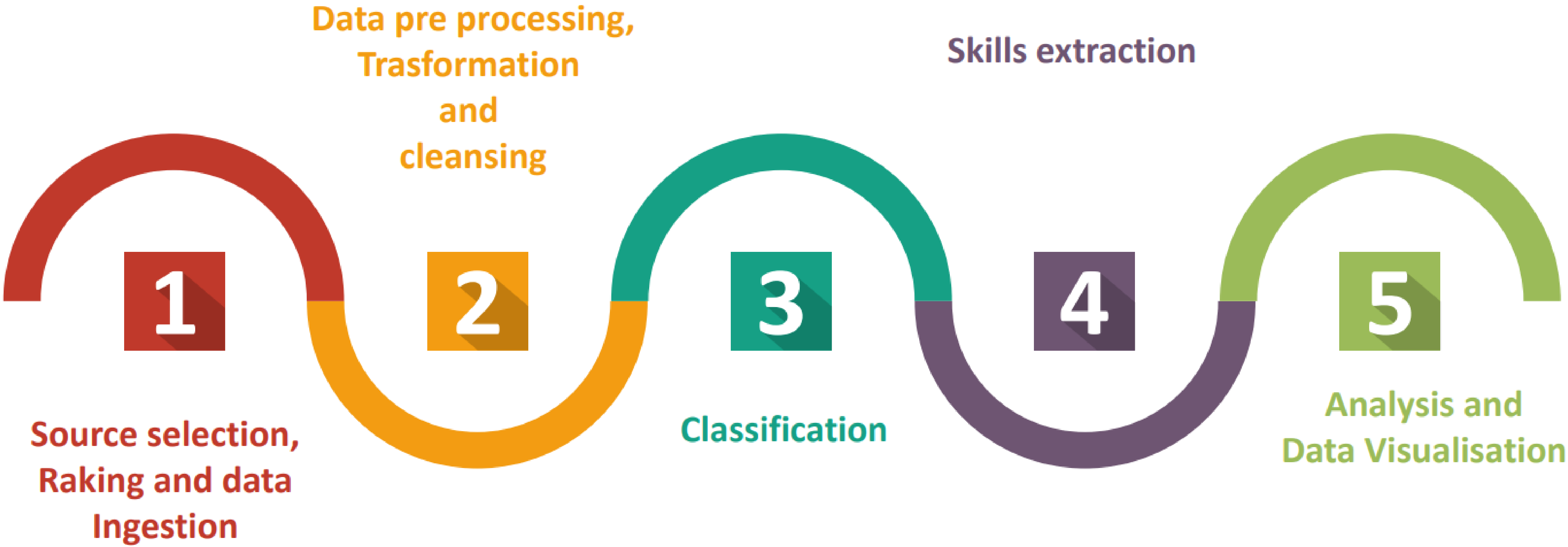
Methodology

LinkedIn public datasets:1) LinkedIn API provides a sanctioned method for accessing platform within terms of use 2) via web scraping without violate LinkedIn's terms of service and data privacy regulations.

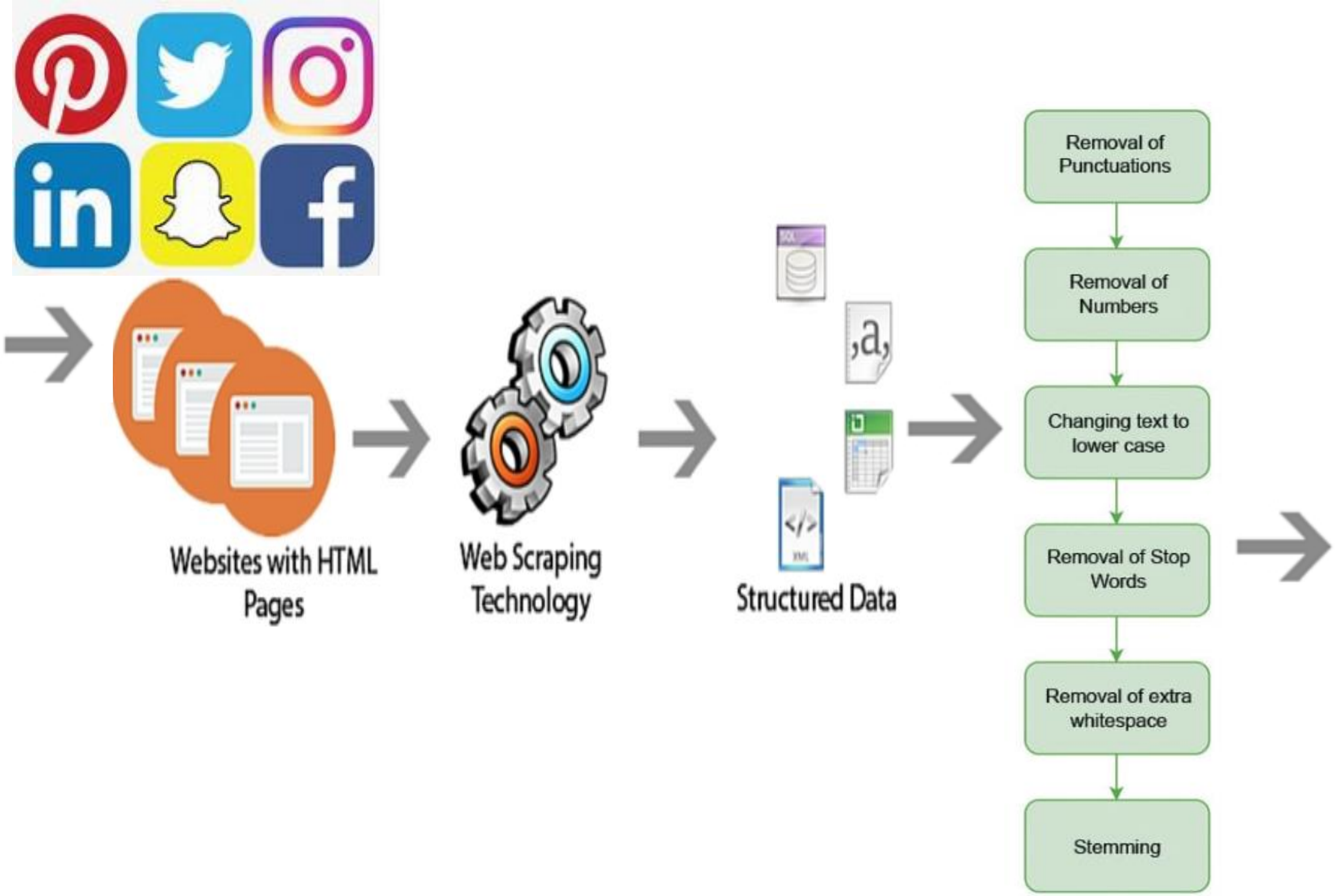
LinkedIn proprietary datasets: Available through LinkedIn's premium products, such as Sales Navigator and LinkedIn Talent Insights. They offer exclusive access to certain data points not available in public datasets(free trial).

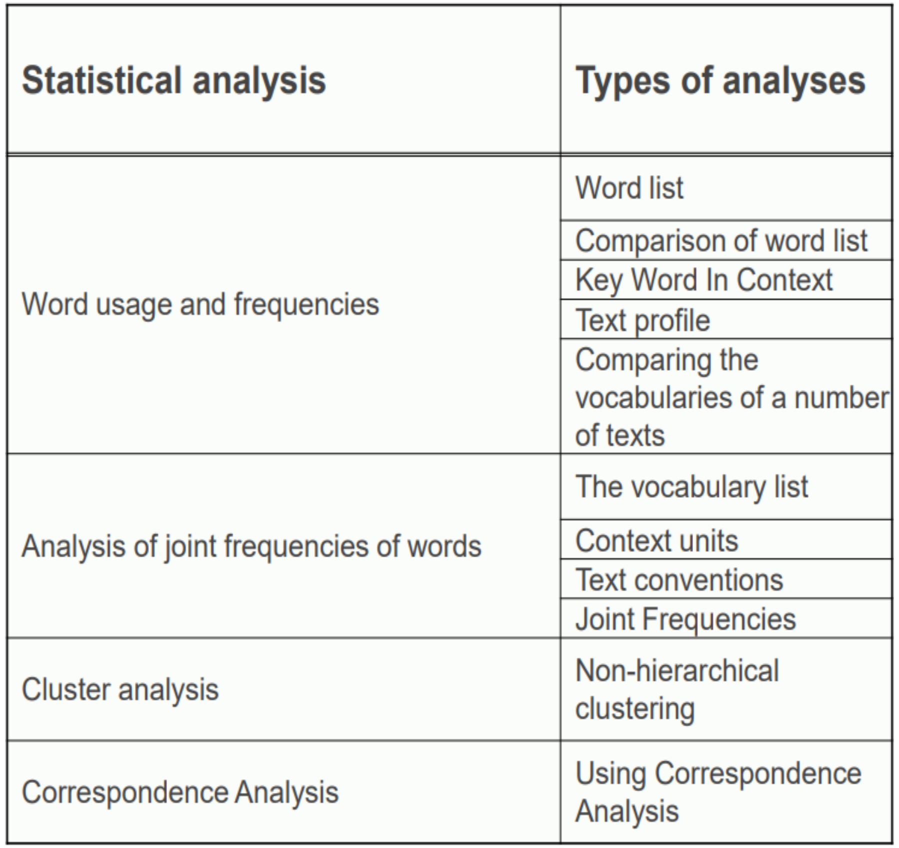
LinkedIn advertising campaign management tool: By simulating a campaign, it is possible to see the estimated number of users on LinkedIn in real time.

The Process


Data pre processing,
Trasformation
and
cleansing
Skills extraction
1
2
3
4
5
Source selection,
Raking and data
Ingestion
Classification
Analysis and
Data Visualisation

LinkedIn as a new source of data





Statistical analysis	Types of analyses
Word usage and frequencies	Word list
	Comparison of word list
	Key Word In Context
	Text profile
	Comparing the vocabularies of a number of texts
Analysis of joint frequencies of words	The vocabulary list
	Context units
	Text conventions
	Joint Frequencies
Cluster analysis	Non-hierarchical clustering
Correspondence Analysis	Using Correspondence Analysis

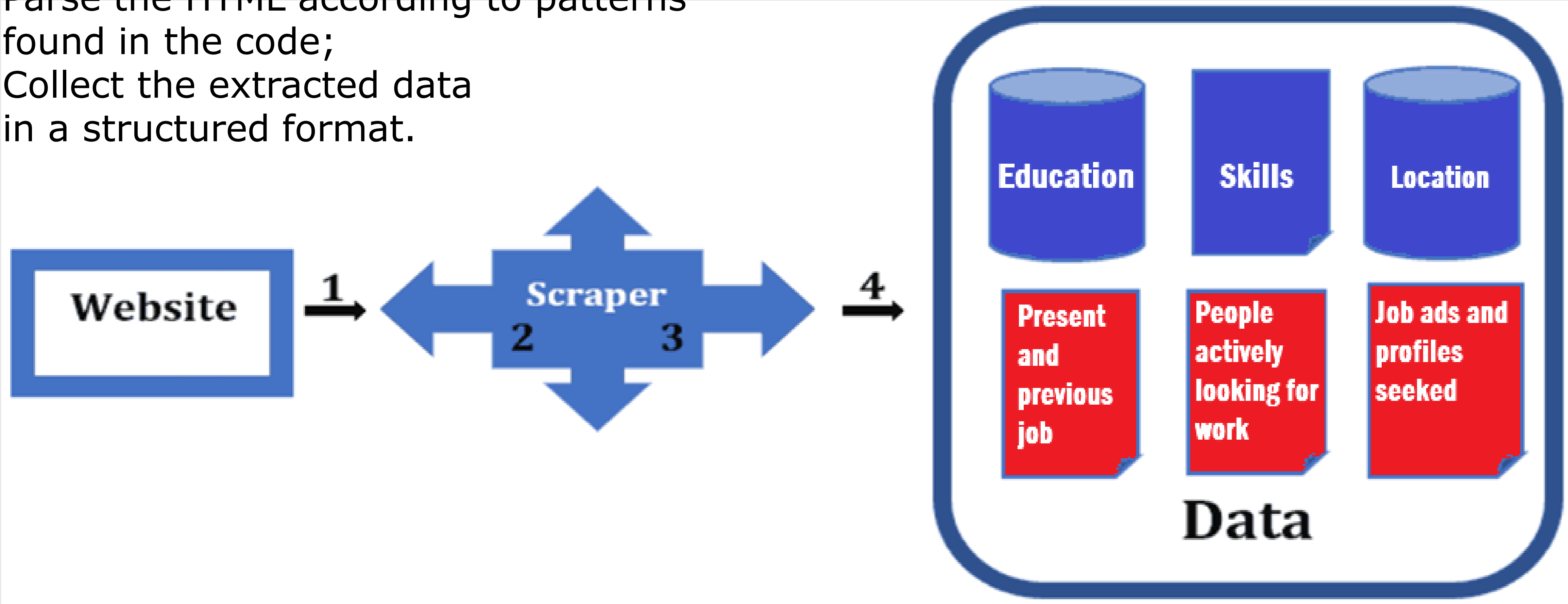
LinkedIn as a new source of data

With A.I.:

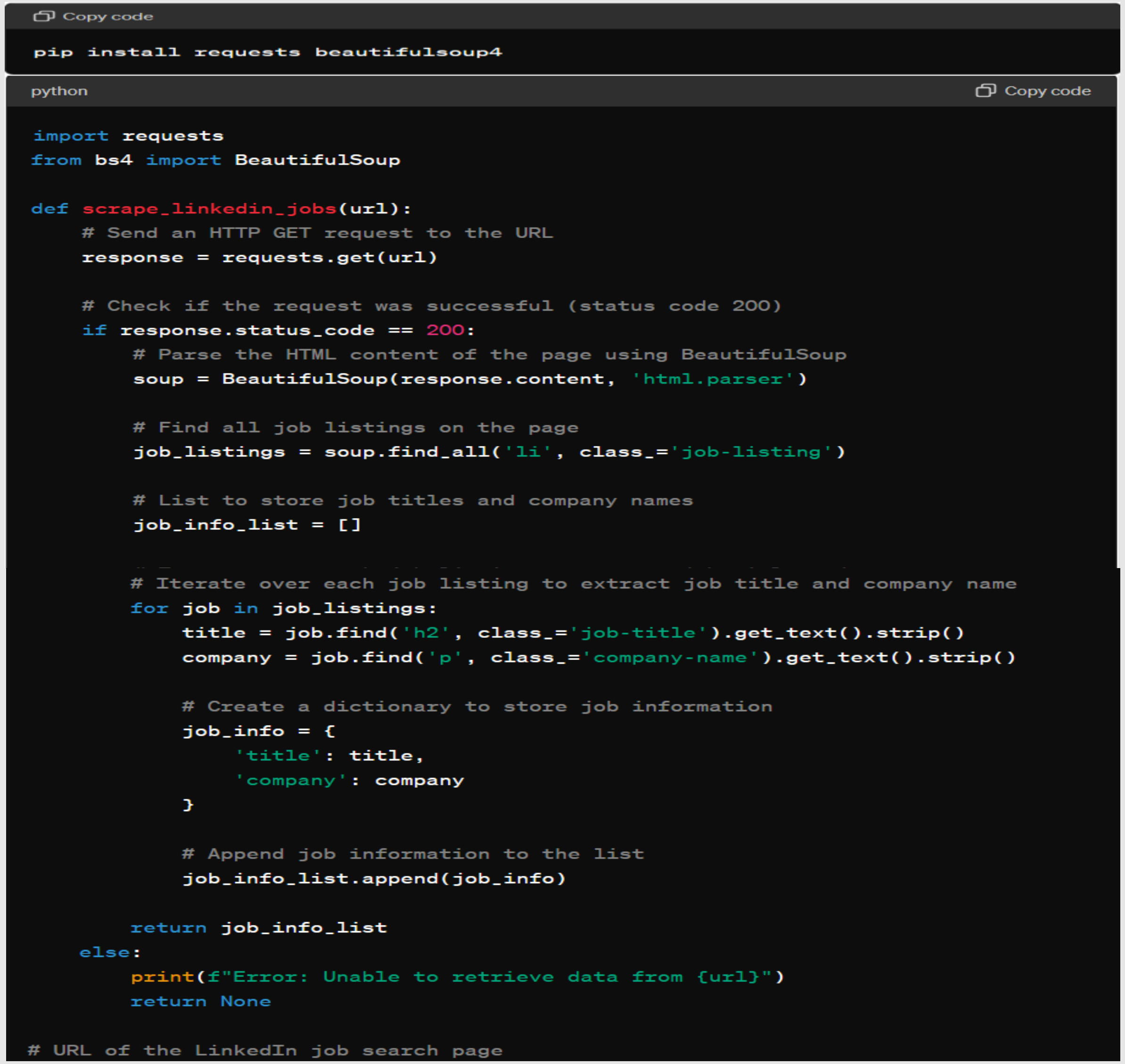
Download web pages

Parse the HTML according to patterns found in the code;

Collect the extracted data in a structured format.



LinkedIn as a new source of data



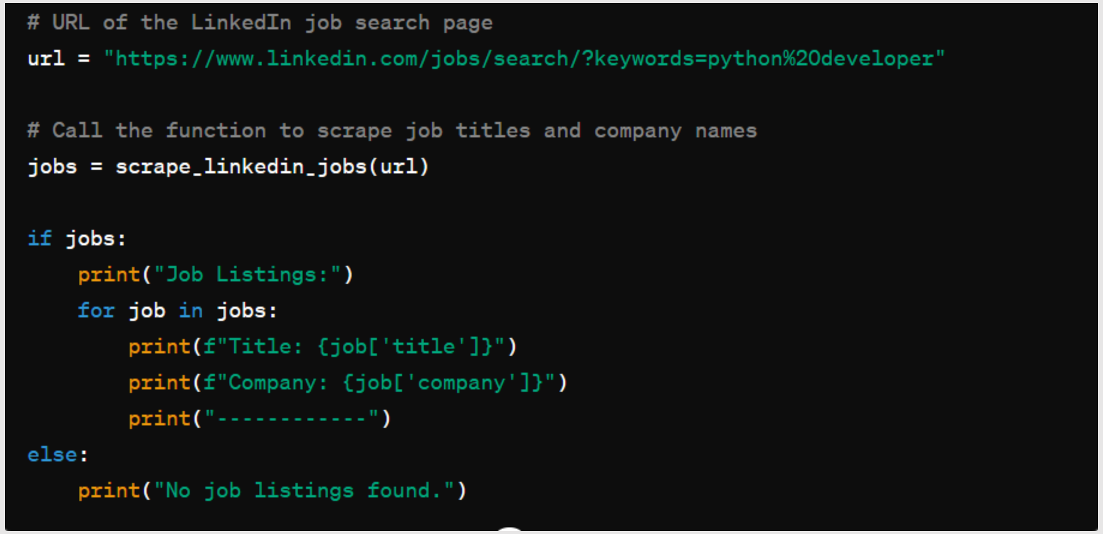
Copy code

```
pip install requests beautifulsoup4
```

python Copy code

```python
import requests
from bs4 import BeautifulSoup

def scrape_linkedin_jobs(url):
    # Send an HTTP GET request to the URL
    response = requests.get(url)

    # Check if the request was successful (status code 200)
    if response.status_code == 200:
        # Parse the HTML content of the page using BeautifulSoup
        soup = BeautifulSoup(response.content, 'html.parser')

        # Find all job listings on the page
        job_listings = soup.find_all('li', class_='job-listing')

        # List to store job titles and company names
        job_info_list = []

        # Iterate over each job listing to extract job title and company name
        for job in job_listings:
            title = job.find('h2', class_='job-title').get_text().strip()
            company = job.find('p', class_='company-name').get_text().strip()

            # Create a dictionary to store job information
            job_info = {
                'title': title,
                'company': company
            }

            # Append job information to the list
            job_info_list.append(job_info)

        return job_info_list
    else:
        print(f"Error: Unable to retrieve data from {url}")
        return None

# URL of the LinkedIn job search page
```

Web scraping without violate LinkedIn's terms of service and data privacy regulations.

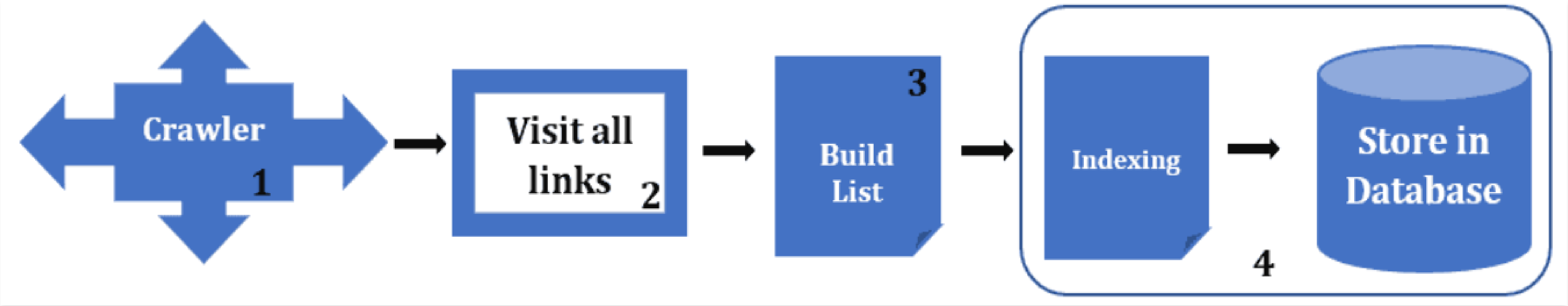


```python
# URL of the LinkedIn job search page
url = "https://www.linkedin.com/jobs/search/?keywords=python%20developer"

# Call the function to scrape job titles and company names
jobs = scrape_linkedin_jobs(url)

if jobs:
    print("Job Listings:")
    for job in jobs:
        print(f"Title: {job['title']}")
        print(f"Company: {job['company']}")
        print("------------")
else:
    print("No job listings found.")
```

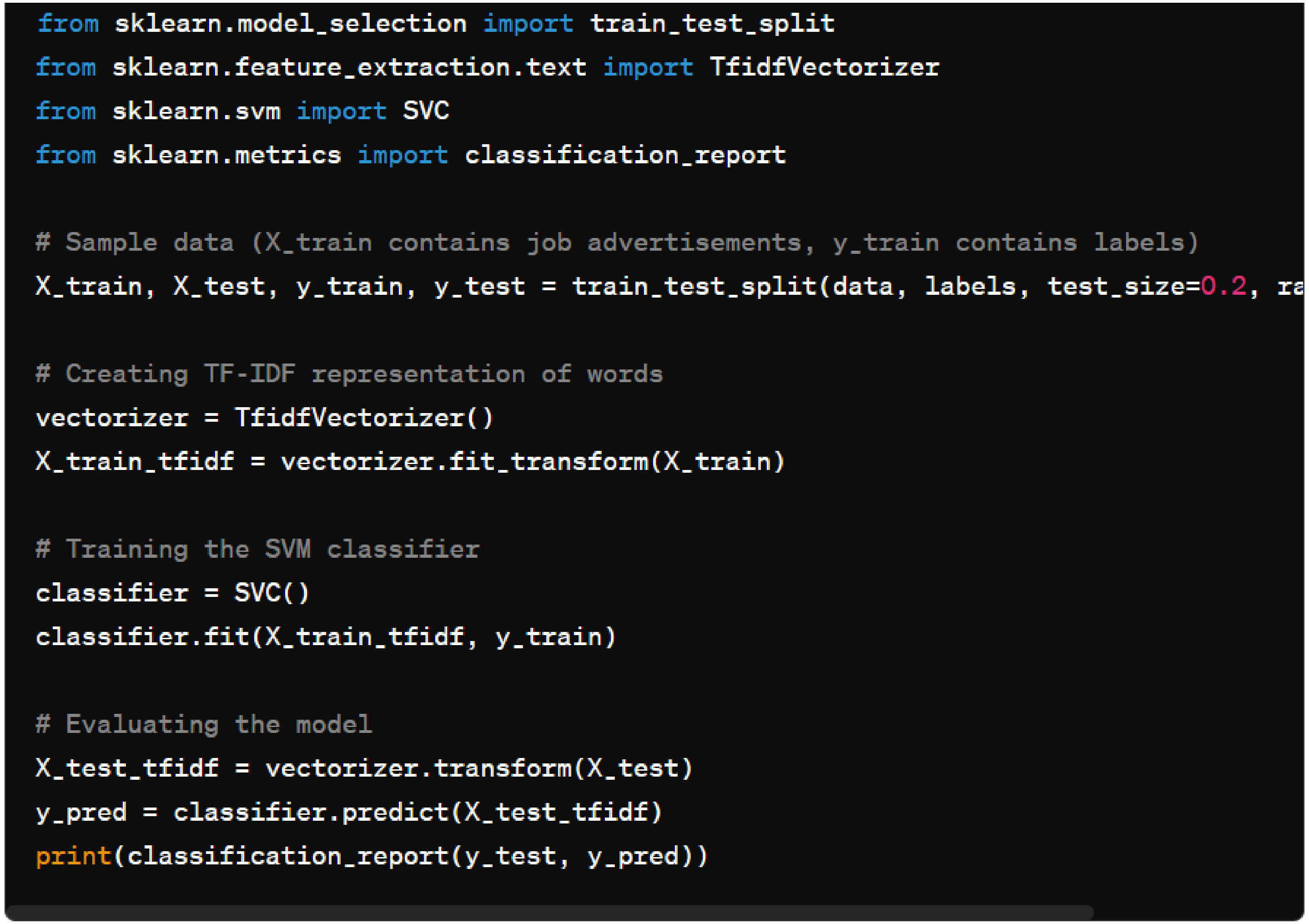
LinkedIn as a new source of data

Web crawling vs Web scraping

1)Take out the given URLs from the crawel frontier
2) Visit each page linked by these URLs
3)Review and classify web pages
4)Index the found URL data and saves it in the database



LinkedIn as a new source of data



```python
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.metrics import classification_report

# Sample data (X_train contains job advertisements, y_train contains labels)
X_train, X_test, y_train, y_test = train_test_split(data, labels, test_size=0.2, ra

# Creating TF-IDF representation of words
vectorizer = TfidfVectorizer()
X_train_tfidf = vectorizer.fit_transform(X_train)

# Training the SVM classifier
classifier = SVC()
classifier.fit(X_train_tfidf, y_train)

# Evaluating the model
X_test_tfidf = vectorizer.transform(X_test)
y_pred = classifier.predict(X_test_tfidf)
print(classification_report(y_test, y_pred))
```

A basic workflow using TF-IDF for feature extraction and an SVM classifier for classification
$TFIDF(t,d)=TF(t,d)IDF(t)$

Methodology

1. Data Preprocessing

•**Tokenization**: Split each profile/ job advertisement into tokens (words or phrases).
•**Removing Punctuation and Stopwords**:
•**Stemming or Lemmatization**:

2. Feature Extraction

•**TF-IDF (Term Frequency-Inverse Document Frequency)**: Calculate the importance of a word relative to a corpus of job advertisements.
•**Word Embeddings**: Represent words as distributed numerical vectors using techniques like Word2Vec or GloVe.

3. Model Definition for Classification

•**Choose a Classification Model**: Use supervised learning algorithms such as Support Vector Machines (SVM), Naive Bayes, Decision Trees, Random Forests, or neural networks like Multi-layer Perceptron (MLP).
•**Training the Model**: Use the labeled dataset (job vs. non-job) to train the classification model.

Methodology

4. Model Evaluation

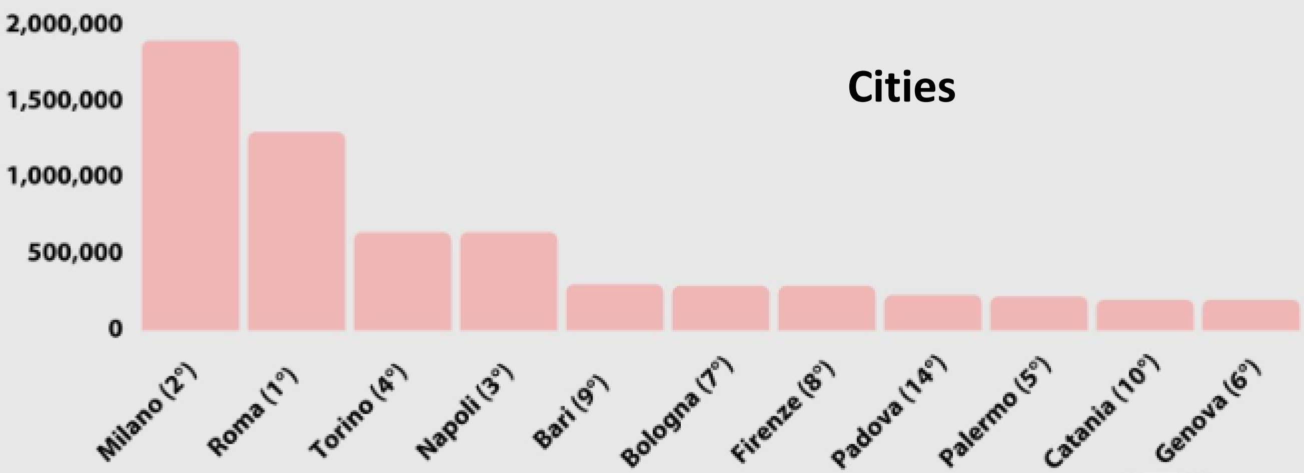
- **Data Splitting**: Split the dataset into training and testing sets.
- **Performance Metrics**: Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1-score.

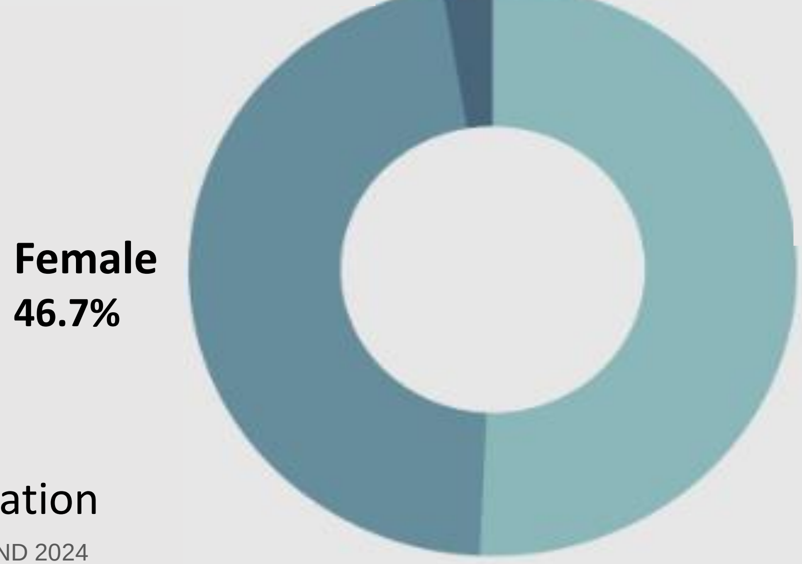
5. Model Optimization and Validation

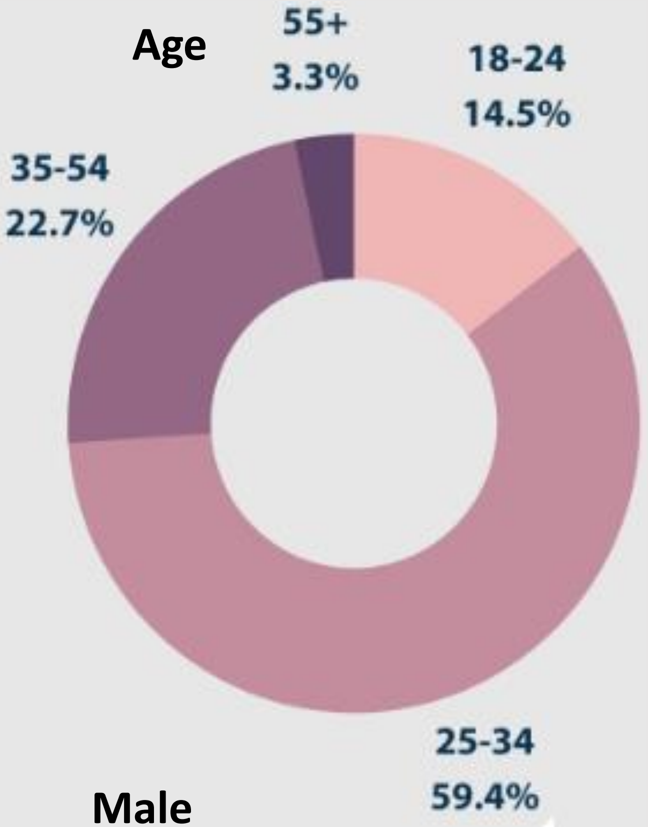
- **Hyperparameter Tuning**: Optimize the model's hyperparameters to improve performance.
- **Cross-Validation**: Validate the model using techniques like k-fold cross-validation to ensure its generalizability.

Linkedin users: Results

Who are Linkedin users in Italy?



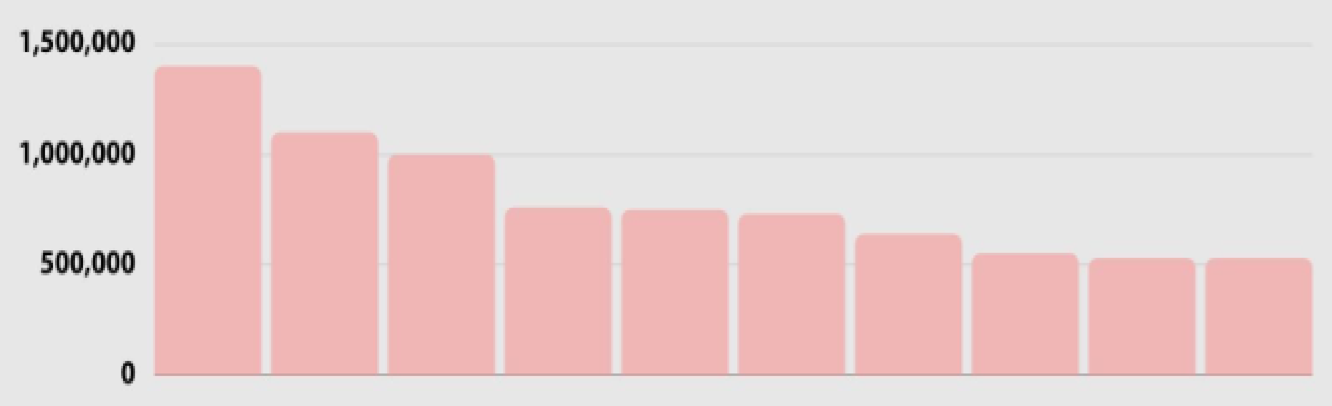




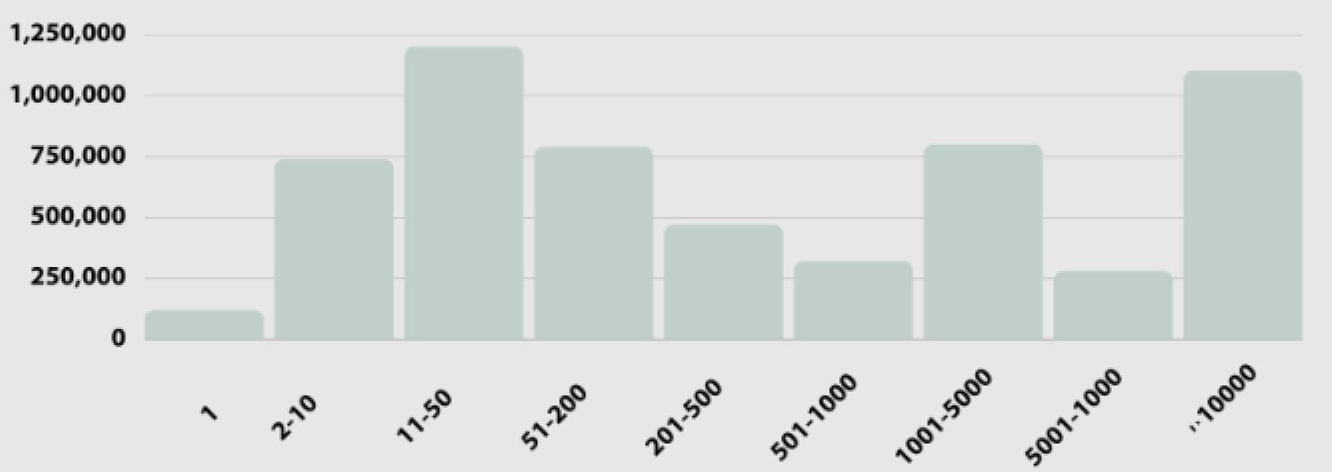
()the position in the ranking of cities by population

LinkedIn Companies in Italy

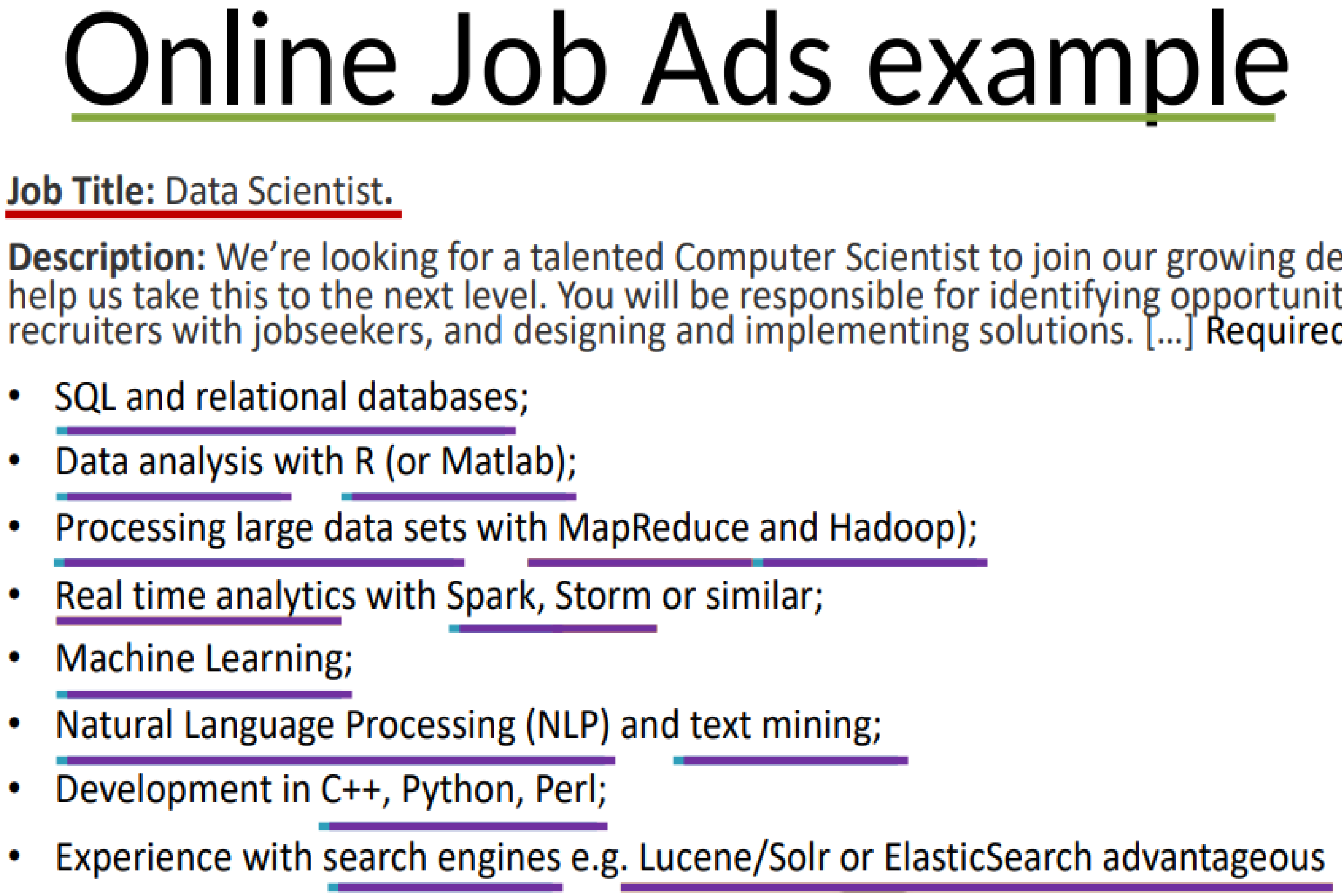
Business sectors



Company size



LinkedIn as a new source of data



Online Job Ads example

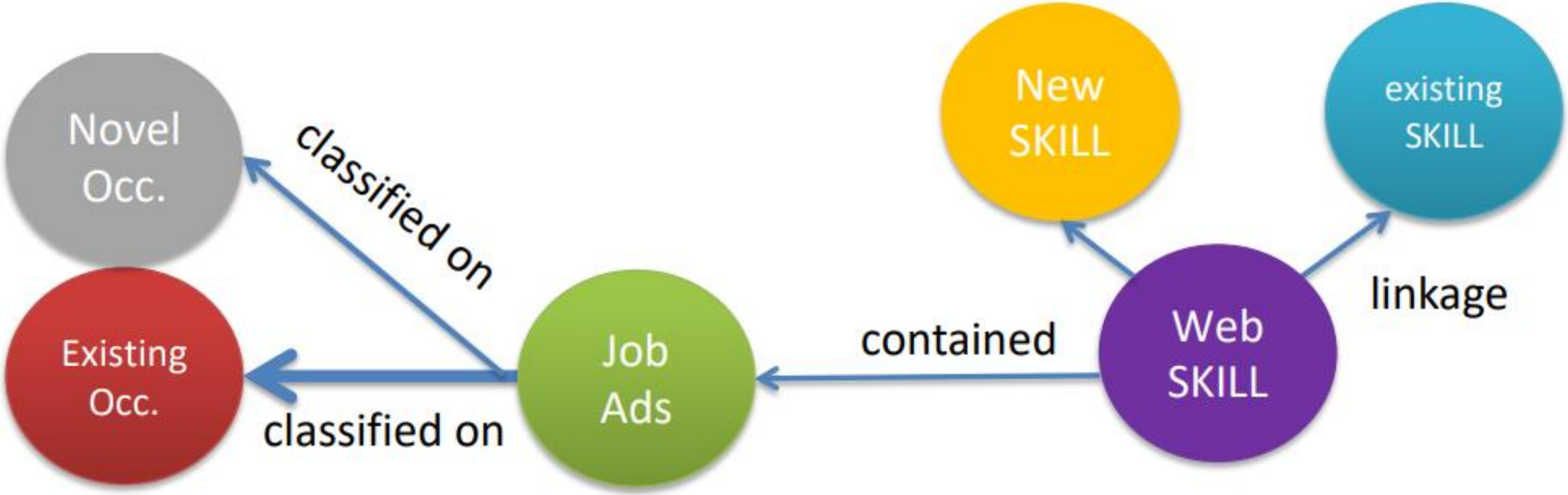
Job Title: Data Scientist.

Description: We're looking for a talented Computer Scientist to join our growing development team. Your expertise in data will help us take this to the next level. You will be responsible for identifying opportunities to further improve how we connect recruiters with jobseekers, and designing and implementing solutions. […] Required skills and experience:

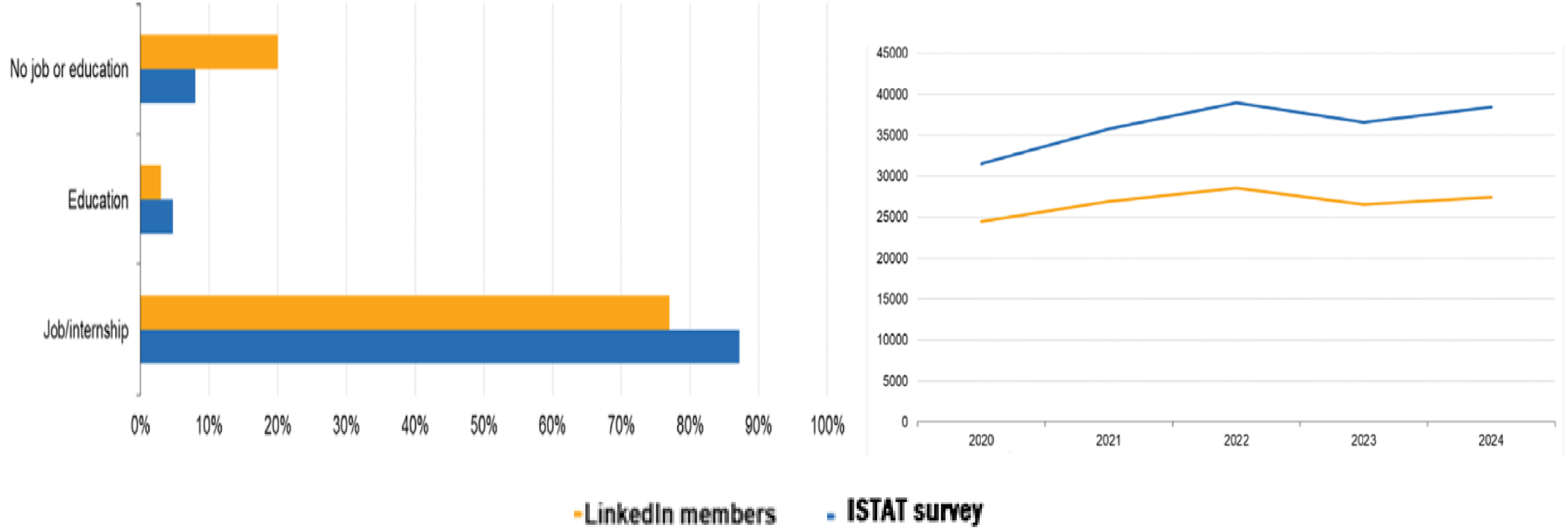
- SQL and relational databases;
- Data analysis with R (or Matlab);
- Processing large data sets with MapReduce and Hadoop);
- Real time analytics with Spark, Storm or similar;
- Machine Learning;
- Natural Language Processing (NLP) and text mining;
- Development in C++, Python, Perl;
- Experience with search engines e.g. Lucene/Solr or ElasticSearch advantageous

1. Data driven approach
2. High granularity
3. High frequency (updated)
4. Focus on important skills

LinkedIn as a new source of data



New data collection results compared traditional survey results



Labour participation of graduated
joung people

LinkedIn as a new source of data: Results

- ✓ We see a structural divergence in the labour participation around 10%, especially for graduates.
- ✓ People may only list jobs on LinkedIn that mach their Education. Labour force survey has data about all the jobs held by an individual.
- ✓ LinkeDin is more popular in some industries than in ithers. That may produce a distorted image as well

Conclusions

01 Much policy demand for information on job openings and the labour market as a whole

02 The new sources may be able to answer questions that traditional statistics cannot

03 Could reduce costs for NIS

04 Could reduce burden for survey respondents,

05 Could improve quality and timeliness of existing statistics

LinkedIn
Connect to Opportunity™

"Aside from the people, the hours, the work, the pay, the stress and the migraines, this is the best job I ever had."

Thank you!

Istat | Istituto Nazionale di Statistica